Entity-Centric Sentiment Classifier for Social Media Analysis

Introduction / Progress

Presented by
Cristobal Leiva
Supervised by
Dr. Simon Scerri
Prof. Dr. Sören Auer

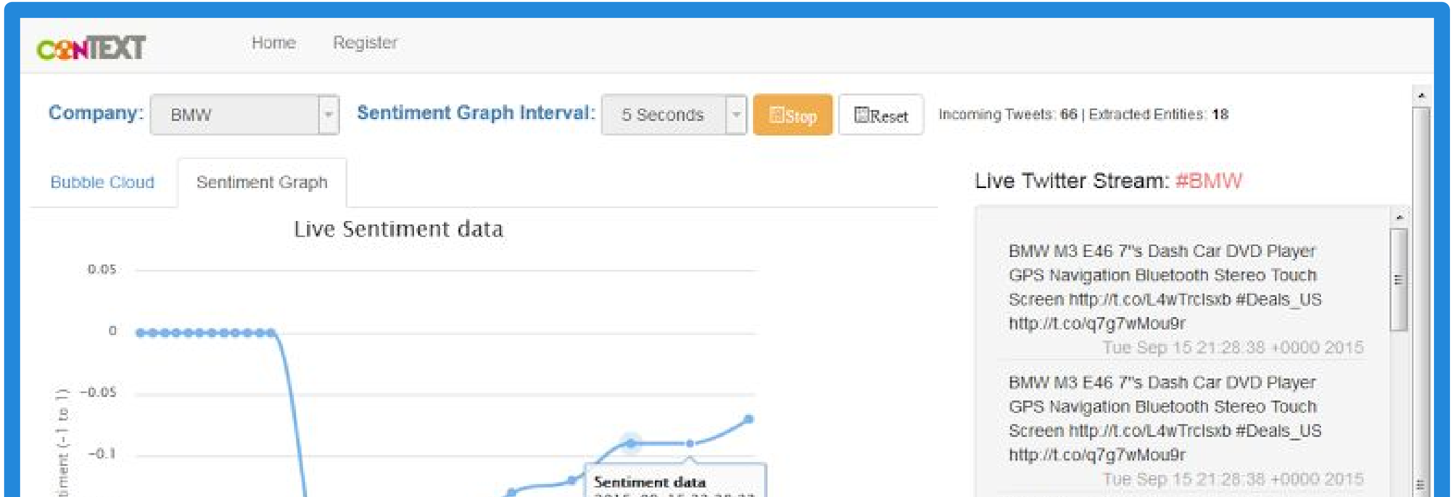
universitätbonn
EIS
ENTERPRISE
INFORMATION SYSTEMS

Motivation

- Social media networking services such as Twitter provide a massive amount of valuable data.
- Core business processes such as market-sensing, customer acquisition and customer relationship management (CRM).
- Cross domain applications: Politics, Sociology and others.

Motivation

- **Linked Data-based Social Media Analysis for Stock Market Tracking.**
  - ReSA (Real-time Sentiment Analysis) by Dr. Ali Khalili.
  - Find correlation between public sentiments and intra-day stock prices.



Company: BMW

Sentiment Graph Interval: 5 Seconds

Incoming Tweets: 66 | Extracted Entities: 18

Live Twitter Stream: #BMW

BMW M3 E46 7"s Dash Car DVD Player GPS Navigation Bluetooth Stereo Touch Screen http://t.co/L4wTrclsxb #Deals_US http://t.co/q7g7wMou9r
Tue Sep 15 21:28:38 +0000 2015

BMW M3 E46 7"s Dash Car DVD Player GPS Navigation Bluetooth Stereo Touch Screen http://t.co/L4wTrclsxb #Deals_US http://t.co/q7g7wMou9r
Tue Sep 15 21:28:38 +0000 2015

Problem

- Determining when a positive or negative sentiment is being expressed along a text span is not enough.
- Real-time analysis environments become a challenge.
- Tweets might contain opinions toward different entities.

Objective

Ultimate goal is to categorize the sentiment towards particular entities in a tweet.

“my iPhone is better than your Nexus 4”

Approach Overview

- Development of a 3 - Class machine learning based sentiment classifier.

Positive - Neutral - Negative

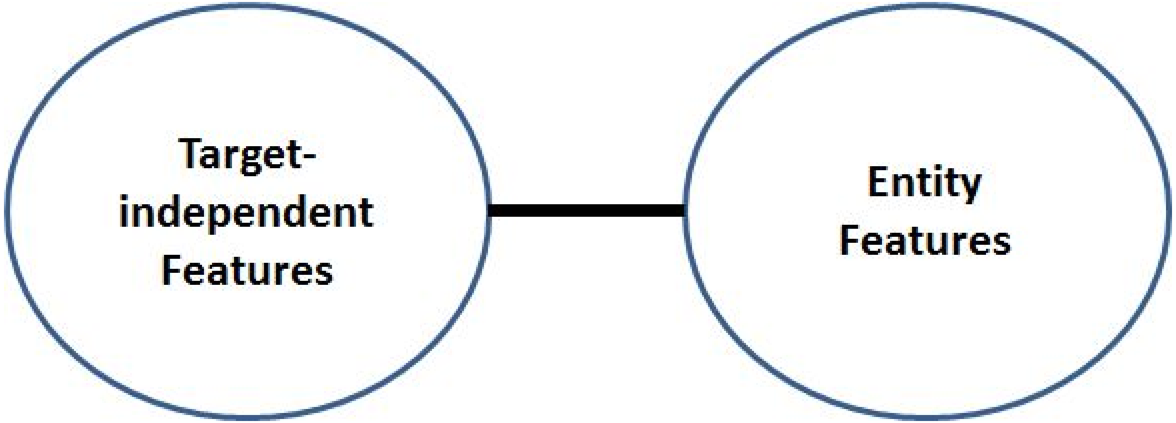
- SVM classifier trained with annotated tweets.
- Inclusion of target-dependent features on the feature-extraction phase relying on Entity-context and natural language rules.

Approach - Preprocessing

- **Tokenizer**
  - Trim text → Unicode Rep. → @ Rep. → URL Removal...
  - Sentence segmentation and stopwords removal.
- **Proprocessor**
  - Slang Correction → Fix Elongation → Negation Context Tagging
- **POS Tagger**
  - Assign part-of-speech (POS) labels to preprocessed tokens.

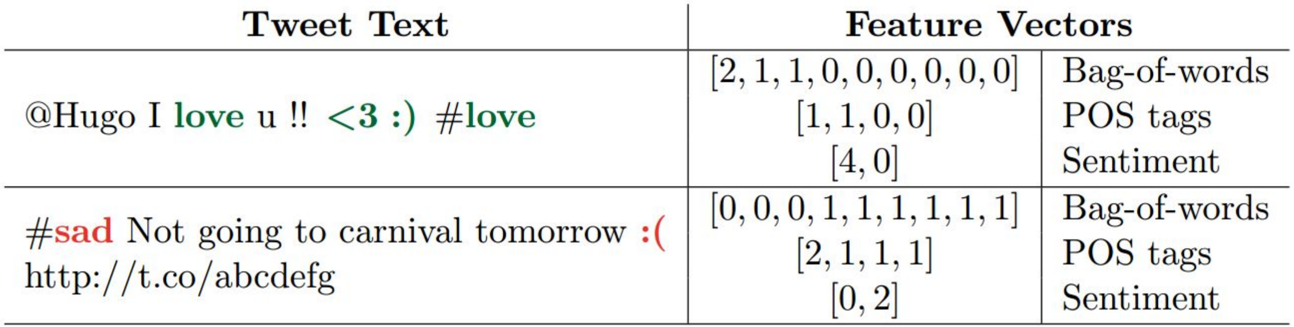
Nouns / *Adjectives* / *Verbs* / *Adverbs* / ...

Approach - Feature Extraction



Approach - Target-independent Features

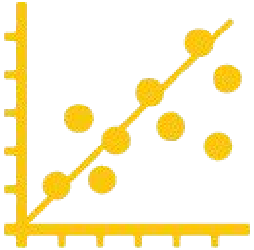
- **Content Features**:
  - # → Neg. Context / all-caps / POS / Hashtags / emoticons / elongated words / exclamation marks / ...
  - Unigrams / bag-of-words model / Boolean term frequency



Tweet Text	Feature Vectors	
@Hugo I love u !! <3 :) #love	[2, 1, 1, 0, 0, 0, 0, 0, 0] [1, 1, 0, 0] [4, 0]	Bag-of-words POS tags Sentiment
#sad Not going to carnival tomorrow :(http://t.co/abcdefg	[0, 0, 0, 1, 1, 1, 1, 1, 1] [2, 1, 1, 1] [0, 2]	Bag-of-words POS tags Sentiment

Approach - Entity Features

- **Named Entity Recognition**:
  - DBpedia Spotlight service for entity annotation.
- **Sentence-Entity features**:
  - # → presence target-entity / sentences without target
- **Entity context Lexicon features**:
  - “But” Clause Rules, NL Rules (“*better than*”)
- **Lexicons**: (7) - Entity context based
  - Manual: AFINN / BingLiu / NRC Emotion Lexicon
  - Semi-Automatic: SentiWordNet / MQPA
  - Automatic: NRC Sentiment140 / Hashtag Lexicon_NOT



Evaluation - Results

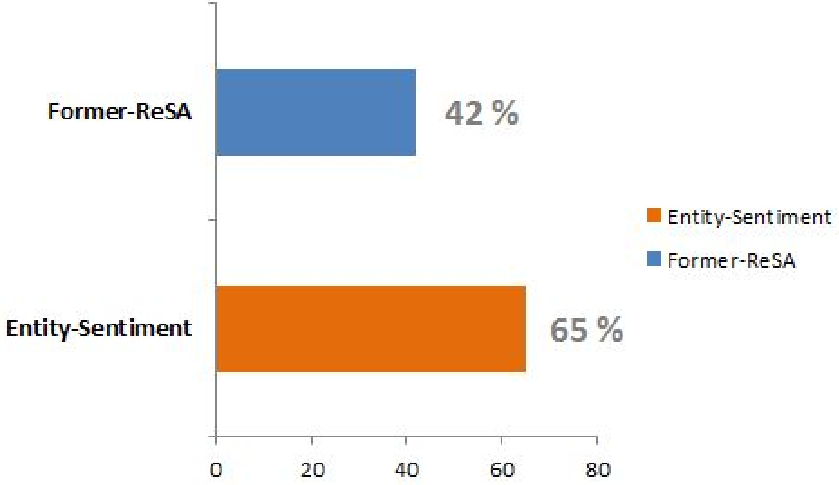
Collection of 4900 Entity-centric annotated tweets.

- Semeval 2015 (Semantic Evaluation) - task 10 - Training data
- Semeval 2016 - task 4 - Training data
- Twitter Sanders Analytics Corpus
- STS - Gold (Saif M. Mohammad)

70% - 30% SVM Training / Eval ratio.

Evaluation - Results (So far)

- **Classification Accuracy:**
  - Number of correct predictions made divided by the total number of predictions made. 4-fold Cross-validation.



Next...

- Evaluation extension
  - Extracted-features evaluation results
  - Evaluate AlchemyAPI.
  - Further testing...

- ReSA SentiTrack Experiment
  - Results and conclusions.

Thank You

Presented by
Cristobal Leiva

Supervised by
Dr. Simon Scerri
Prof. Dr. Sören Auer